Health data poverty: an assailable barrier to equitable digital health care

Hussein Ibrahim, Xiaoxuan Liu, Nevine Zariffa, Andrew D Morris, Alastair K Denniston**

Data-driven digital health technologies have the power to transform health care. If these tools could be sustainably delivered at scale, they might have the potential to provide everyone, everywhere, with equitable access to expert-level care, narrowing the global health and wellbeing gap. Conversely, it is highly possible that these transformative technologies could exacerbate existing health-care inequalities instead. In this Viewpoint, we describe the problem of health data poverty: the inability for individuals, groups, or populations to benefit from a discovery or innovation due to a scarcity of data that are adequately representative. We assert that health data poverty is a threat to global health that could prevent the benefits of data-driven digital health technologies from being more widely realised and might even lead to them causing harm. We argue that the time to act is now to avoid creating a digital health divide that exacerbates existing health-care inequalities and to ensure that no one is left behind in the digital era.

Lancet Digit Health 2021

Published Online
March 4, 2021
https://doi.org/10.1016/S2589-7500(20)30317-4

*Joint senior authors

Centre for Regulatory Science and Innovation, Birmingham Health Partners, Birmingham, UK (H Ibrahim MBChB, X Liu MBChB, Prof A K Denniston PhD); **University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK** (H Ibrahim, X Liu, Prof A K Denniston); **Academic Unit of Ophthalmology, Institute of Inflammation and Ageing, College of Medical and Dental Sciences, University of Birmingham, Birmingham, UK** (H Ibrahim, X Liu, Prof A K Denniston); **Health Data Research UK, London, UK** (X Liu, Prof A K Denniston, Prof A D Morris PhD); **NMD Group, Bala Cynwyd, PA, USA** (N Zariffa MMath); **NIHR Biomedical Research Centre at Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, UK** (Prof A K Denniston)

Correspondence to:
Prof Alastair K Denniston, Academic Unit of Ophthalmology, Institute of Inflammation and Ageing, College of Medical and Dental Sciences, University of Birmingham, Birmingham B15 2TT, UK
a.denniston@bham.ac.uk

Introduction

Developments in artificial intelligence, and particularly machine learning, have shown the power that data-driven digital health technologies have to transform health care.[1] If these tools could be sustainably delivered at scale, they have the potential to provide everyone, everywhere, with equitable access to expert-level care—in line with the vision set out in WHO's Global Strategy on Digital Health[2]—and thereby narrow the global health and wellbeing gap. Conversely, it is highly possible that these transformative technologies could exacerbate existing health-care inequalities instead.[3–8]

In this Viewpoint, we describe the problem of health data poverty. We then discuss potential solutions to the problem, notably investing in inclusive and representative health datasets to support equitable discovery and innovation in digital health care.

The problem: health data poverty and its risk of creating a digital health divide

We define health data poverty as the inability for individuals, groups, or populations to benefit from a discovery or innovation due to insufficient data that are adequately representative. To properly understand the problem of health data poverty, and, by extension, the solutions, it is important to understand what health data are, how health data are used, and what health data disparities are.

First, what are health data? Health data have been defined as information relating to the past, current, or future physical or mental health status of a person.[9] Health data include any of the clinical, biochemical, radiological, molecular, and pathological information pertaining to a patient that is captured by health-care professionals and, increasingly, digitally recorded and stored in electronic patient health records.[10] Health data also include any of the health-related information that is captured by patients themselves, by using sensors and a growing range of smart devices, and stored in smartphones, tablets, computers, and cloud-based repositories.[10] The digitalisation of health care has meant that humans today are amassing health data at astronomical rates, with estimates putting the total volume of health data in 2020 at 2314 exabytes—where one exabyte is equal to 1 billion gigabytes.[11]

Second, how are health data used? Health data are used to benefit patients and the public, and the use of health data can be classified as either primary or secondary. The primary (direct) use is where health data are used to deliver health care to the individual from whom they were collected.[12] Here, health data inform health-care professionals when making diagnoses and decisions relating to an individual's care. The secondary (indirect) use is where health data are used to improve health care and health-care services for a population.[12] In this instance, health data relating to many different people are pooled together to create large health datasets that can be used to evaluate multiple hypotheses related to improving health care. Appropriate interrogation of health data can improve understanding of disease and disability, identify better ways to predict and diagnose illness, develop new treatments and technologies, monitor safety, plan services, and evaluate policies. The secondary use of health data has been instrumental in driving discovery and innovation in the digital age. As the medical community attempts to shift from reactive care to proactive and preventive care, there is a move to integrate predictive devices into daily life. This process is contributing to a future of health care that is more preventive, predictive, and personalised, and it is this use that is most relevant to health data poverty.

Third, what are health data disparities? Health data disparities are systematic differences in the quantity or quality, or both, of health data representing different individuals, groups, or populations. Data disparities have been shown within and between populations as well as across different demographics, disciplines, and diseases. For example, as of 2018, individuals included in genome-wide association studies were 78% European, 10% Asian, 2% African, 1% Hispanic, and less than 1% all other ethnic groups, leading to the conclusion that European bias in human genetic studies is "both scientifically damaging and unfair".[13] Similarly, the sample of

individuals whose data make up UK Biobank has been shown to be much healthier than the general population, with underrepresentation of individuals with socioeconomic deprivation and from particular ethnic backgrounds.[14] However, disparities are not solely limited to genetics datasets. Regarding imaging, it was found that, across multiple medical specialities in the USA, deep learning algorithms that perform image-based diagnostic tasks were disproportionately trained on data from California, Massachusetts, and New York, with little to no representation from the remaining 47 states and their populations.[15] And, focusing just on the field of ophthalmology, a global review of ophthalmic imaging datasets revealed disparities in the representation of different populations and disease groups in publicly available health data repositories.[16]

These imbalances lead to datasets that underrepresent key segments of the overall population. As these same datasets are used to develop and validate digital health technologies, a possible extreme scenario is that data-driven interventions are safe and effective for some people, but dangerous and ineffective for others. This scenario is possible because these tools are entirely dependent on the data that are used to develop and validate them and can be highly sensitive to fundamental characteristics including age, sex, ethnicity, environment, and potentially many others. Therefore, people who are underrepresented might be unable to benefit from these data-driven interventions, and could even be harmed by them. This is a well recognised and common cause of the limited usability of some artificial intelligence and machine learning algorithms beyond health care. For example, this is the reason why automatic speech recognition systems such as Alexa and Siri can have high failure rates when applied to diverse user groups.[17]

In cases in which a data-driven digital health technology underperforms when applied to individuals from an underrepresented group, individuals from the underrepresented group can be considered to be health data poor. This lack of generalisability of discoveries and innovations across peoples and populations due to underrepresentation is also rife in non-digital contexts. It is remarkable how frequently tests, therapeutics, and other interventions are not evaluated in children,[18] women (especially pregnant women),[19,20] people from minority ethnic groups,[21] and older people,[22] and indeed, this situation has recently been highlighted by the US Food and Drug Administration in their guidance for diversity in clinical trials.[23] Similarly, it is surprising how infrequently the so-called normal ranges of physiological parameters take into account age, sex, and ethnicity.[24] But it is in the accelerating world of digital health that health data poverty will cause most direct and indirect harm, and that is our focus in this Viewpoint.

Evidence is already accumulating of the negative consequences of health data poverty in the context of digital health. In 2019, Tomašev and colleagues[25] described a deep learning algorithm for predicting acute kidney injury in adults. Only 6·4% of the data in the training dataset was derived from female patients and, as expected, the model was found to underperform in this group. This is a good example of gender-related health data disparities causing relative health data poverty in women. The authors acknowledge that "validating the predictive performance of the proposed system on a general population would require training and evaluating the model on additional representative datasets".[25] The impact of health data poverty is only going to increase as more data-driven technologies become mainstream. An area of intense interest in digital health is automated skin lesion diagnosis, but there are already concerns about the failure of these digital health applications to be inclusive, due in part to the non-representativeness of the training data.[26] It should, however, be noted that the failures of digital health in this regard mimic long-standing failures of humans, with provision of training materials for dermatologists being overwhelmingly from paler skin types,[27] and with human diagnostic performance being consistently worse in darker skinned patients.[28] Although examples such as these might be relatively obvious, there could be other applications in which the underperformance of data-driven technologies in particular poorly represented groups will not be detected unless this is specifically looked for.

It is encouraging that the risks associated with the use of non-representative datasets,[4,5] as well as the associated ethical issues that can arise throughout the development pipeline of a digital health technology,[3] are being increasingly recognised. The value of the term health data poverty is to recognise this data paucity for what it is: a poverty that is born out of existing inequalities and could foster further inequality. There is a risk that, rather than narrowing the health and wellbeing gap around the world, data-driven solutions, such as artificial intelligence systems, will create a digital health divide that leaves us with a two-tier health system of digital haves and digital have-nots. The digital divide cannot be thought of as simply a matter of access to technologies, but rather as an issue that runs through the whole digital pathway, starting with the data foundation on which these technologies are dependent. If health-care systems wish to embrace the advantages of digital health without perpetuating or exacerbating existing health inequalities, they must recognise and address the threat of health data poverty.

Addressing health data poverty: where do we start?

As with other forms of poverty, health data poverty is complex and not amenable to a single, simple solution. Our purpose here is to raise awareness of this important issue, to highlight areas of progress, and to outline some

 Published online March 4, 2021 https://doi.org/10.1016/S2589-7500(20)30317-4

broad areas that should be considered if we are to address this issue.

Increasing awareness within data and digital health communities, and beyond

There is an opportunity for the health data research and digital health communities to be advocates for data-deprived individuals, groups, and populations, so as to ensure that they are not left behind.[3–5] Ultimately, funders, regulators, policy makers, and politicians need to make it a requirement for creators of digital health solutions to provide assurance that these technologies will be able to perform across different populations and settings. It is down to the health data research and digital health communities to ensure that these funders, regulators, policy makers, and politicians are educated in this regard. Efforts such as the Journal of the American Medical Informatics Association's special focus issue on health informatics and health equity, published in 2019, are important steps in the right direction.[29]

Initiatives to increase transparency as to the composition of training datasets in artificial intelligence and machine learning models, such as the Data Nutrition Project and Datasheets for Datasets, could increase awareness and accountability and promote more responsible practice within data and digital health communities.[30,31] Similarly, efforts to encourage people developing these models to transparently assess and report how representative the training dataset is of the model's intended population and how well the model performs across relevant subgroups within that population, could facilitate the safe and appropriate application of such models in the real world.[32,33] This information can be included in model facts labels or model cards, tools designed to communicate essential information about individual models to the user.

Transparent, effective communication to citizens

Transparently and effectively communicating to citizens as to how their data can directly contribute to making digital health solutions more safe, effective, and equitable is of utmost importance.[34] Studies, including deliberative research and citizens' juries, show that the more that people understand about how their data are used in health data research, the more they are willing to participate, by sharing their data, in such research.[35–37] Additionally, it is essential to alleviate citizens' privacy and confidentiality concerns,[34] which have been shown to be substantial barriers that prevent people from participating in health data research.[35–37] Principally, alleviation of these concerns will be achieved by ensuring that all research activities are underpinned by effective governance systems, but also by transparently and effectively communicating to citizens the ways in which their data are being used and protected.[38]

Improving equity of digital access for data-gathering as well as health provision

An increasing proportion of digital health solutions are direct-to-citizen through smartphones, wearables, and other devices. There is, however, much inequity of access to these devices due to a range of factors including economic, education, and social. In the UK, in 2018, the Office for National Statistics reported that 10% of adults were internet non-users; this was associated with being older, female, from minority ethnic groups, and disabled.[39] At the same time, the Lloyds Bank UK Consumer Index estimated that 8% of people in the UK did not have basic digital skills and a further 12% had only limited abilities.[40] It is worth noting that there is the belief, among already-excluded groups, that they do not need to engage digitally; for example, 64% of households without internet access in the UK say that the reason they do not have internet access is that they do not need it.[39] The impact that this paucity of digital access and engagement can have has been highlighted by the COVID-19 pandemic, during which digital health care has become central to the delivery of many services; a large number of people could be excluded from these digital health-care services.[41]

This paucity of digital access among a significant proportion of people is also relevant to health data poverty. Data from smartphone apps and wearables (the use of which is low among those who are not digitally engaged) will increase the scale and breadth of health data records to levels beyond the records kept by conventional clinical systems. These sources will increasingly provide the data that train future data-driven technologies. A recent example is the effect of symptom-reporting smartphone apps such as the COVID Symptom Study app (ZOE). With over 4 million users worldwide, this app has been an exceptional real-time data-gathering exercise, supporting novel insights such as clustering of COVID-19 symptoms and the prevalence of self-reported anosmia as a symptom.[42] Crucially, the app also highlights the boundaries of generalisability outside the digital population, such as underrepresentation of older people who have a much higher prevalence of anosmia for reasons not related to COVID-19.[43]

A more developed understanding is needed of the digital determinants of health; that is to say, the way in which social, cultural, and economic factors influence access to and outcomes of digital health solutions.[44] Only through a thorough understanding of these factors will it be possible to take appropriate and effective action to ensure equitable access and outcomes regardless of age, sex, ethnicity, education, income, and geography.

Building inclusive and representative datasets to support equitable discovery and innovation in digital health care

Building datasets is probably the least exciting aspect of, but most important foundation for, digital health. We assert that the development and validation of digital health

solutions, especially those involving artificial intelligence and machine learning systems, requires investment in datasets that are: sufficiently representative of the whole population into which they will be deployed; of sufficient quantity and quality to provide confidence in any external validation process (and retraining if needed); and of appropriate accessibility, recognising the contrasting needs of development and training (high accessibility to anonymised data within a safe system)[45] versus independent evaluation and regulation (highly restricted access to data that is not available to developers but only to regulators and those providing independent testing and assurance of those algorithms). This requirement for datasets satisfying these criteria applies both to newly created digital health systems and when considering the deployment of an established system into a new setting or population; assurance of performance cannot be assumed on the basis of previous results in different settings or populations.

For **COVID-19 Therapeutics Accelerator** see https://www.therapeuticsaccelerator.org

The creation of such datasets is a large investment. In the UK, Health Data Research UK is committed to "uniting the UK's health data to enable discoveries that improve people's lives". It was recognised quickly that this aim could not be achieved simply through the use of research cohorts consisting of people who opt in to having their health data collected for the purposes of research. Such cohorts are valuable but lead to skewed populations, as seen in the example of UK Biobank.[14] More representative datasets are instead gathered through the use of routinely collected National Health Service data, such as through a number of Health Data Research Hubs and from Data Controllers who are working together as part of the Health Data Research Alliance, both of which are initiatives convened by Health Data Research UK.[46] During the COVID-19 pandemic, this access to routinely collected data has meant that the UK has been well placed to address how the disease has affected different sectors of the population, including highlighting the worse outcomes seen in Black and Asian ethnic groups, and exploring the underlying causes for these findings. In other industries, datasets from a broad range of participants and environments are made publicly available to maximise accuracy of artificial intelligence and machine learning innovations, for example speech recognition and natural language processing datasets.[47,48]

For **Health Data Research UK** see https://www.hdruk.ac.uk/

For **Infectious Diseases Data Observatory** see https://www.iddo.org

For **Health Data Research Alliance** see https://ukhealthdata.org

The focus here, however, is not on what has been achieved in the UK or other countries with relatively mature health data resources, but rather on how much is still to do, and the concern that many vulnerable groups and whole parts of the world are being left behind. This will require concerted international data sharing initiatives as a powerful mechanism to address COVID-19 and other global challenges. Among these is the International COVID-19 Data Research Alliance and Workbench,[49] supported by the COVID-19 Therapeutics Accelerator, which has the vision to unite data from international clinical trials, biomedical research, and health research to enable discoveries that benefit all people, everywhere, by reducing the harm of the COVID-19 pandemic. This builds on the work of the Infectious Diseases Data Observatory and other international initiatives that assemble clinical, laboratory, and epidemiological data on a collaborative platform to be shared with the research and humanitarian communities. These are important first steps in developing and maintaining the integrity of a trustworthy international health data ecosystem.

Various philanthropic organisations fund international research initiatives. Data generated from these investments could be contracted to be shared, and targets for the adequate representation of the intended population be imposed as a condition of the grant; the funder could also dictate the level of open access required. By doing this, funders will enhance the original aim of the research, as well as the onward use of the resulting data in training datasets to better represent all people.

Search strategy and selection criteria

References for this Viewpoint were identified through searches of PubMed, Google Scholar, and Google search engine, with the search terms "artificial intelligence", "big data", "digital divide", "digital health", "digital health equity", "equity", "health disparities", "health equity", "health informatics", and "machine learning", from date of database inception until Nov 24, 2020. Searches were supplemented by manually screening the references of relevant articles. Only papers published in English language were reviewed. The final reference list was generated on the basis of originality and relevance to this Viewpoint.

Looking to the future, together

Addressing health data poverty requires a collective approach of international stakeholders to enable the necessary engagement and investment, to share learning, and, wherever possible, to operate to common goals and standards. Solutions will vary between countries according to their needs, resources, and type of health system, but will need to consider core issues, such as ensuring that all research activities are underpinned by effective governance systems. This long-term investment in and stewardship of data requires a different mandate to the short-term drivers that predominate in academic, industry, health service, or political sectors. Funding models may vary but should ensure that data are available for the benefit of the population from which they were collected, with the opportunity for ongoing development and testing of digital health technologies that will improve the health of that population. Prioritisation of datasets will also vary by country but factors that should be considered will include local health needs, amenability of those needs to digital health solutions, and the facility

and cost of acquiring the data needed to support that digital health solution.

As the Ada Lovelace Institute reminds us: "Missing data matters: it can exacerbate inequalities on a societal scale. When that data is operationalised into algorithmic decision-making systems and AI, the social processes that produce racial inequality—mechanisms of power, economics, knowledge, culture and language—can be written into technologies with huge societal impacts."[50] The barrier of missing data is assailable; now is the time to act if we are to counter the emerging digital health divide and build the data infrastructure that means that all parts of society can benefit from digital health solutions.

Contributors

HI and AKD prepared the manuscript. XL, NZ, and ADM revised the manuscript. All authors read and approved the final manuscript.

Declaration of interests

NZ reports personal fees from the Bill & Melinda Gates Foundation, AstraZeneca, Genentech, Bristol Myers Squibb, and Anova Enterprises outside the submitted work; stock equity in AstraZeneca, GlaxoSmithKline, Johnson & Johnson, Merck, Moderna, Pfizer, Sanofi, Takeda, TranslateBio, Vaxart, Vir Biotechnology, and Inovio Pharmaceuticals; and stock options in Anova Enterprises. NZ was employed by AstraZeneca from 2011 to 2019. All other authors declare no competing interests.

Acknowledgments

The views expressed in this article are those of the authors and not necessarily those of the National Institute for Health Research, or the Department of Health and Social Care.

References

1 Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019; **25:** 44–56.
2 WHO. Global strategy on digital health 2020–2025. https://www.who.int/docs/default-source/documents/gs4dhdaa2a9f352b0445bafbc79ca799dce4d.pdf?sfvrsn=f112ede5_50 (accessed Nov 6, 2020).
3 Chen IY, Pierson E, Rose S, Joshi S, Ferryman K, Ghassemi M. Ethical machine learning in health care. *ArXiv* 2020; published online Oct 8. https://arXiv:2009.10576 (preprint).
4 Veinot TC, Mitchell H, Ancker JS. Good intentions are not enough: how informatics interventions can worsen inequality. *J Am Med Inform Assoc* 2018; **25:** 1080–88.
5 Lee EWJ, Viswanath K. Big data in context: addressing the twin perils of data absenteeism and chauvinism in the context of health disparities research. *J Med Internet Res* 2020; **22:** e16377.
6 Ferryman K, Winn RA. Artificial intelligence can entrench disparities—here's what we must do. Nov 16, 2018. The Cancer Letter. https://cancerletter.com/articles/20181116_1 (accessed Nov 6, 2020).
7 Wiens J, Saria S, Sendak M, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nat Med* 2019; **25:** 1337–40.
8 Ghassemi M, Naumann T, Schulam P, Beam AL, Chen IY, Ranganath R. Practical guidance on artificial intelligence for health-care data. *Lancet Digit Health* 2019; **1:** e157–59.
9 GDPR.EU. Recital 35: Health data. https://gdpr.eu/recital-35-health-data/ (accessed Nov 6, 2020).
10 Data Saves Lives. What is health data? https://datasaveslives.eu/what-is-health-data (accessed Nov 6, 2020).
11 EMC Digital Universe with Research and Analysis by IDC. The digital universe of opportunities: rich data and the increasing value of the internet of things. April, 2014. https://www.emc.com/leadership/digital-universe/2014iview/index.htm (accessed Nov 18, 2020).
12 Safran C, Bloomrosen M, Hammond WE, et al. Toward a national framework for the secondary use of health data: an American Medical Informatics Association white paper. *J Am Med Inform Assoc* 2007; **14:** 1–9.
13 Sirugo G, Williams SM, Tishkoff SA. The missing diversity in human genetic studies. *Cell* 2019; **177:** 26–31.
14 Fry A, Littlejohns TJ, Sudlow C, et al. Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. *Am J Epidemiol* 2017; **186:** 1026–34.
15 Kaushal A, Altman R, Langlotz C. Geographic distribution of US cohorts used to train deep learning algorithms. *JAMA* 2020; **324:** 1212–13.
16 Khan SM, Liu X, Nath S, et al. A global review of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalisability. *Lancet Digit Health* 2020; **3:** E51–66.
17 Sheng LMA, Edmund MWX. Deep learning approach to accent classification. Stanford University Project Report 2017. http://cs229.stanford.edu/proj2017/final-reports/5244230.pdf (accessed Nov 18, 2020).
18 Fernandez C, Canadian Paediatric Society, Bioethics Committee. Ethical issues in health research in children. *Peadiatr Child Health* 2008; **13:** 707–12.
19 Holdcroft A. Gender bias in research: how does it affect evidence based medicine? *J R Soc Med* 2007; **100:** 2–3.
20 Blehar MC, Spong C, Grady C, Goldkind SF, Sahin L, Clayton JA. Enrolling pregnant women: issues in clinical research. *Womens Health Issues* 2013; **23:** e39–45.
21 Redwood S, Gill PS. Under-representation of minority ethnic groups in research—call for action. *Br J Gen Pract* 2013; **63:** 342–43.
22 Witham MD, McMurdo MET. How to get older people included in clinical studies. *Drugs Aging* 2007; **24:** 187–96.
23 U.S. Food and Drug Administration. Guidance document: enhancing the diversity of clinical trial populations – eligibility criteria, enrollment practices, and trial designs. November, 2020. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/enhancing-diversity-clinical-trial-populations-eligibility-criteria-enrollment-practices-and-trial (accessed Nov 20, 2020).
24 Whyte MB, Kelly P. The normal range: it is not normal and it is not a range. *Postgrad Med J* 2018; **94:** 613–16.
25 Tomašev N, Glorot X, Rae JW, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature* 2019; **572:** 116–19.
26 Adamson AS, Smith A. Machine learning and health care disparities in dermatology. *JAMA Dermatol* 2018; **154:** 1247–48.
27 Buster KJ, Stevens EI, Elmets CA. Dermatologic health disparities. *Dermatol Clin* 2012; **30:** 53–59.
28 Dawes SM, Tsai S, Gittleman H, Barnholtz-Sloan JS, Bordeaux JS. Racial disparities in melanoma survival. *J Am Acad Dermatol* 2016; **75:** 983–91.
29 Veinot TC, Ancker JS, Bakken S. Health informatics and health equity: improving our reach and impact. *J Am Med Inform Assoc* 2019; **26:** 689–95.
30 Holland S, Hosny A, Newman S, Joseph J, Chmielinski K. The dataset nutrition label: a framework to drive higher data quality standards. *ArXiv* 2018; published online May 9. https://arXiv:1805.03677 (preprint).
31 Gebru T, Morgenstern J, Vecchione B, et al. Datasheets for datasets. *ArXiv* 2020; published online Mar 19. https://arXiv:1803.09010 (preprint).
32 Sendak MP, Gao M, Brajer N, Balu S. Presenting machine learning model information to clinical end users with model facts labels. *NPJ Digit Med* 2020; **3:** 41.
33 Mitchell M, Wu S, Zaldivar A, et al. Model cards for model reporting. *ArXiv* 2019; published online Jan 14. https://arXiv:1810.03993v2 (preprint).
34 Institute of Medicine. Beyond the HIPAA privacy rule: enhancing privacy, improving health through research. Washington, DC: The National Academies Press, 2009.
35 Kalkman S, van Delden J, Banerjee A, Tyl B, Mostert M, van Thiel G. Patients' and public views and attitudes towards the sharing of health data for research: a narrative review of the empirical evidence. *J Med Ethics* 2019; published online Nov 12. https://doi.org/10.1136/medethics-2019-105651.
36 Sheridan R, Martin-Kerry J, Hudson J, Parker A, Bower P, Knapp P. Why do patients take part in research? An overview of systematic reviews of psychosocial barriers and facilitators. *Trials* 2020; **21:** 259.

37 OneLondon, Ipsos MORI, The King's Fund. Public deliberation in the use of health and care data. https://www.onelondon.online/wp-content/uploads/2020/07/Public-deliberation-in-the-use-of-health-and-care-data.pdf (accessed Nov 18, 2020).

38 Mello MM, Lieou V, Goodman SN. Clinical trial participants' view of the risks and benefits of data sharing. *N Engl J Med* 2018; **378:** 2201–12.

39 Office for National Statistics. Exploring the UK's digital divide. https://www.ons.gov.uk/peoplepopulationandcommunity/householdcharacteristics/homeinternetandsocialmediausage/articles/exploringtheuksdigitaldivide/2019-03-04 (accessed Nov 6, 2020).

40 Lloyds Bank. Lloyds Bank UK consumer digital index 2020. https://www.lloydsbank.com/assets/media/pdfs/banking_with_us/whats-happening/lb-consumer-digital-index-2020-report.pdf (accessed Nov 6, 2020).

41 Majeed A, Maile EJ, Coronini-Cronberg S. Covid-19 is magnifying the digital divide. Sept 1, 2020. The BMJ Opinion. https://blogs.bmj.com/bmj/2020/09/01/covid-19-is-magnifying-the-digital-divide/?utm_campaign=shareaholic&utm_medium=twitter&utm_source=socialnetwork (accessed Nov 6, 2020).

42 Menni C, Sudre CH, Steves CJ, Ourselin S, Spector TD. Quantifying additional COVID-19 symptoms will save lives. *Lancet* 2020; **395:** e107–08.

43 Menni C, Sudre CH, Steves CJ, Ourselin S, Spector TD. Widespread smell testing for COVID-19 has limited application – authors' reply. *Lancet* 2020; **396:** 1630–31.

44 Crawford A, Serhal E. Digital health equity and COVID-19: the innovation curve cannot reinforce the social gradient of health. *J Med Internet Res* 2020; **22:** e19361.

45 University of the West of England. Five safes: designing data access for research. https://www2.uwe.ac.uk/faculties/bbs/Documents/1601.pdf (accessed Nov 6, 2020).

46 Health Data Research UK. Our hubs. https://www.hdruk.ac.uk/help-with-your-data/our-hubs-across-the-uk/ (accessed Nov 6, 2020).

47 Lionbridge. 20 best speech recognition datasets for machine learning. https://lionbridge.ai/datasets/best-speech-recognition-datasets-for-machine-learning/ (accessed Nov 18, 2020).

48 Open Data Science. 20 open datasets for natural language processing. https://medium.com/@ODSC/20-open-datasets-for-natural-language-processing-538fbfaf8e38 (accessed Nov 18, 2020).

49 Health Data Research UK. International COVID-19 data research alliance and workbench. https://www.hdruk.ac.uk/covid-19/international-covid-19-data-alliance/ (accessed Nov 6, 2020).

50 Ada Lovelace Institute. Black data matters: how missing data undermines equitable societies. https://www.adalovelaceinstitute.org/black-data-matters-how-missing-data-undermines-equitable-societies/ (accessed Nov 6, 2020).